INCORPORATING SUPPLEMENTAL SAMPLE DATA INTO A NATIONAL SURVEY

Dwight K. French, Energy Information Administration
Marcus J. Sanchez and Dwight B. Brock, National Center for Health Statistics

INTRODUCTION

The primary responsibility of the National Center for Health Statistics is the development and maintenance of survey mechanisms that provide accurate and comprehensive information on matters of health, health resources, vital events, and related matters. To satisfy the health community's needs for data pertaining to nursing homes, residents and staff of nursing homes, and related finances and costs, the NCHS conducts the National Nursing Home Survey (NNHS).

The NNHS utilizes a two-stage, stratified, probability sample design. In the first stage, facilities are grouped into 24 strata according to bed-size and type of care. A systematic sample of facilities is selected using an independent random start with a fixed skip interval for each stratum.

After the original national sample for the 1977 survey was drawn, state nursing home estimates were requested by California, Illinois, Massachusetts, New York and Texas. But the NNHS is not designed to support state estimates. After some reflection, it was decided that the national sample contained enough California facilities to make estimates for that State. For each of the other four states a supplemental sample of facilities was drawn in order to accumulate sufficient data to make state estimates (See Table 1).

PROBLEM

The question we are investigating is: Can data from these supplemental facilities be incorporated with the national sample in such a way that national estimates of nursing home parameters are improved?

The difference in stratum formation between the national sample and the supplemental samples is important in the exercise. In the first stage of the national sample, after the facilities have been grouped into 24 strata according to bedsize and type of care, a systematic sample of facilities was selected using an independent random start and a fixed skip interval for each stratum. Thus for the national sample, the stratum information was bedsize-type of care across States. On the other hand, the supplemental samples of facilities were drawn systematically, using an independent random start and a fixed sampling interval for each sampling stratum within each supplemented State. So for the supplemental samples, the stratum formation was bedsize-type of care within State.

This difference in stratum formation (bedsize-type of care strata across States for the national sample, as opposed to bedsize-type of care strata within State for the supplemental samples) creates difficulties in trying to incorporate the supplemental data into a national estimator. Three approaches to this problem were considered:

(1) Incorporate the data from supplemental facilities in stratum i with the data from all facilities in the national sample from stratum i. This approach forces the supplemental facilities to represent facilities in all states, which is contrary to the design of the supplemental samples. Since the supplemental samples encompassed only four States, such a procedure would probably introduce serious bias into the data. An additional, lesser problem is that the weights for all facilities would have to be changed to accommodate this approach.

(2) Reweight the national and supplemental data so that separate stratum i estimates are produced for each State, and then summed to produce national estimates for stratum i. This approach is faithful to the design of the supplemental samples, but not the original national sample. In addition, many state-stratum classes would have to be generated from the combined national and supplemental facilities in the appropriate stratum. Again, serious biases could be expected, and the weights for all facilities would have to be changed.

(3) Incorporate the data from supplemental facilities in State s, stratum i with the data from the national sample from State s, stratum i. Then reweight that data from State s so that the combined national and supplemental facilities from State s represent the same proportion of the facilities in stratum i as the national sample did originally. This approach follows the designs of both the national and supplemental samples more closely than either (1) or (2) as mentioned before. There is bias in the procedure, but only to the extent that a supplemental State is over- or under-represented in the national sample of facilities from stratum i. However, the percentage of U.S. facilities affected by this "estimator bias" is much smaller for this approach than for (1) or (2) as mentioned before. In addition, only the sample facilities in the four supplemented States need to be reweighted as a result of this approach. The facilities in all other States retain their original weight based on the national sample. Because of these advantages, the third approach was used to derive the alternative estimators used in this study.

NOTATION

Let X' = the estimate of aggregate parameter X for all in-scope facilities in the NNHS universe.

X_i' = the estimate of X for all in-scope facilities in stratum i of the NNHS universe.

X_{ij} = the observed value of facility characteristic X for sample facility j in stratum i.

X_{ijk} = the observed value of within-facility characteristic X for sample unit k in sample facility j in stratum i.

X_{sijk} = the observed value of within-facility characteristic X for sample unit k in sample facility j in stratum i in State S.

M_i = the number of facilities in stratum i of the NNHS universe.

m_i = the number of facilities in stratum i that were selected for the national sample.

m_i' = the number of sample facilities from stratum i that were in scope at the time of the survey.

$M_i \cdot \left(\frac{m_i'}{m_i}\right) = M_i'$ = the estimated number of in-scope facilities in stratum i of the NNHS universe.

$\dot{m}_i$ = the number of in-scope facilities in the national sample that responded to the facility identification questionnaire. ($\dot{m}_i \leq m_i$)

m_{1si}, m_{1si}', $\dot{m}_{1si}$ = the analogues of m_i, m_i' and $\dot{m}_i$ for State s within stratum i of the national sample.

m_{2si} = the number of facilities in the supplemental sample from State s, stratum i.

m_{2si}' = the number of facilities in the supplemental sample from State s, stratum i that were in-scope at the time of the supplemental survey.

$\dot{m}_{2si}$ = the number of in-scope facilities in the supplemental sample from State s, stratum i that responded to the facility identification questionnaire.

B_{ij} = the number of beds in facility j of stratum i.

$B_i = \sum_{j=1}^{M_i} B_{ij} = B_i$ = the number of beds in all facilities in stratum i of the NNHS universe (in or out of scope).

$B_i' = \frac{M_i}{m_i} \sum_{j=1}^{m_i} B_{ij} = B_i'$ = the estimated number of beds in all facilities in stratum i (in or out of scope) based on the national sample.

N_{sij} = the total number of within-facility sampling units in facility j of stratum i, State s

N_i' = the total number of within-facility sampling units in stratum i for facilities in the national sample for which a sampling unit list was submitted.

$\dot{N}_i$ = the total number of within-facility sampling units in stratum i for facilities in the national sample for which the sampling unit list was submitted and one or more sampling unit questionnaires were completed.

$\dot{n}_{sij}$ = the number of within-facility sampling units for which questionnaires were completed for facility j in stratum i, State s.

$\dot{m}_{ix}$ = the number of in-scope sample facilities in the national sample of stratum i that responded to all questions necessary to estimate characteristic X. (Facilities that responded to the facility identification questionnaire might not have had other data forms completed, such as the expense questionnaire or the resident or staff sampling lists).

The subscript i in the preceding notation represents the sampling stratum. Sixteen of these strata were common to the national and supplemental samples. These 16 strata were formed by crossing the dichotomous type-of-care variable with eight bedsize categories, for facilities whose type of care was known prior to the survey (see Table 1). An additional eight strata were used in the national design but were not included in any of the supplemental samples. These strata were defined by the same eight bedsize categories for facilities whose type of care was unknown prior to the survey.

The basic stratum estimator for the 1977 NNHS is

$$X'_i = \frac{M'_i}{\dot{m}_{ix}} \sum_{j=1}^{\dot{m}_{ix}} X_{ij}$$

The following kinds of estimators were proposed to incorporate supplemental sample data (in the formulae $\sum_s^S$ denotes summation over the four supplemented States, $\sum_s^{Ns}$ denotes summation over all other States: α_{si}, α_{1si} and α_{2si} are real numbers, $0 \leq \alpha_{si}, \alpha_{1si}, \alpha_{2si} \leq 1$)

The approach used to develop values of the coefficients was to adapt the modeling scheme proposed by Schaible (1978). The estimators are given below.

1. $$X'_{1i} = \frac{M'_i}{\dot{m}_{ix}} \sum_s^{Ns} \sum_{j=1}^{\dot{m}_{1si}} X_{sij} +$$

$$+ \frac{M'_i}{\dot{m}_{ix}} \sum_s^S \frac{\dot{m}_{1si}}{\dot{m}_{1si} + \dot{m}_{2si}} \cdot \left[\sum_{j=1}^{\dot{m}_{1si}} X_{1sij} + \sum_{j=1}^{\dot{m}_{2si}} X_{2sij} \right]$$

for facility characteristics, where

X_{1sij} = observed value of facility characteristic X for national sample facility j in stratum i of state S.

X_{2sij} = observed value of facility characteristic X for supplemental sample facility j in stratum i of state S.

2. $$X'_{2i} = \frac{M'_i}{\dot{m}_{ix}} \sum_s^{Ns} \sum_{j=1}^{\dot{m}_{1si}} X_{sij} +$$

$$+ \sum_s^S \left\{ \alpha_{si} \left[\frac{M'_i}{\dot{m}_{ix}} \sum_{j=1}^{\dot{m}_{1si}} X_{1sij} \right] + (1 - \alpha_{si}) \left[\sum_{j=1}^{\dot{m}_{1si}} X_{1sij} + \right. \right.$$

$$\left. \left. + \frac{M'_i / \dot{m}_{ix} (\dot{m}_{1si}) - \dot{m}_{1si}}{\dot{m}_{2si}} \sum_{j=1}^{\dot{m}_{2si}} X_{2sij} \right] \right\}$$

where $\alpha_{si} = \frac{1}{2}$ for all s and i.

3. X'_{3i} is the same as 2 with the exception that $\alpha_{si} = \frac{\dot{m}_{1si}}{\dot{m}_{1si} + \dot{m}_{2si}}$

Notice that the first equation is the basic stratum estimator or the original estimator for the survey. Also notice the original estimator is the first term in each of the three alternative estimators numbered one, two and three. It is applied to the facilities in non-supplemented states, reflecting the condition that facilities in non-supplemented states retain their original weight based on the national sample. The remaining portions of each of the three equations pertain to facilities in the four supplemented states.

The first alternative estimator (Equation 1) adjusts for the supplemental data by "deflating" the aggregate sum over the $\dot{m}_{1si} + \dot{m}_{2si}$ sample observations in State s to an expected sum over the original $\dot{m}_{1si}$ values before inflating to the State's representation in stratum i.

The first component (premultiplied by α_{si}) of the second and third estimators is the original estimator for stratum i applied to State s. The second component (premultiplied by $1-\alpha_{si}$) treats the $\dot{m}_{1si}$ original observations as self-representing and then inflates the $\dot{m}_{2si}$ supplemental observations to represent the remainder of State s's contribution to the total national estimate for stratum i. The difference between the second and third estimators is in the definition of the α_{si} (see equations 2 and 3).

RESULTS

To evaluate the alternative estimators we produced 220 estimates of selected population parameters for each of the three estimators and compared them with the corresponding estimates based on the original national estimator and the original sample of facilities. The next step was to produce a variance estimate for each alternative estimate of each statistic, and compare them to the variance estimates of the corresponding estimates based on the original sample.

If at least one of the alternative estimators results in substantially lower variance estimates than those based only on the national sample, we can consider ourselves successful in incorporating the supplemental facilities into the national estimator provided there is no evidence of a large bias associated with the procedure. Such success could have implication for future incorporation of small area subsamples into larger area surveys. Otherwise, the original sample would appear to be best for computing national estimates, at least for this survey.

To measure the relative increase or decrease in variance between estimator 1 and the original estimator, the following computation was performed over each of the 220 observations. The same computation was performed for estimators 2 and 3, also. An illustration of the results for a sub-

set of the 220 variables is given in Table 2, along with the average relative difference (ARD) between the alternative and original estimates and their variances.

$$ARD = \frac{1}{220}\left\{\sum_{1}^{220}\left[\frac{\text{Alternative Var.} - \text{Original Var.}}{\text{Original Var.}}\right]\right\}$$

ARD 1 = - 0.0039
ARD 2 = 0.0386
ARD 3 = - 0.0027

Estimator one shows the greatest decrease in variance. Estimator 3 shows the second best. If it were necessary to choose among the new estimators, Estimator 1 appears to be the best.

The results shown here are far less than what one would need in order to recommend the use of alternative estimators 1, 2, or 3. There are undoubtedly other estimators which will perform better than these. In fact, we are currently investigating two others about which we hope to report in the future. Furthermore, since there are but four states for which supplemental information is available, it is unlikely that a great deal more improvement would be achieved under the best of conditions. The important point, however, is that it is becoming more common to see small geographic areas "piggybacking" surveys run by larger government agencies. With constraints on the federal (and other) budgets growing tighter each year, it is quite likely that this practice ("piggybacking") will continue, perhaps to an even greater degree. Thus we need to continue the kinds of investigations described here in order to make the most efficient use of the samples that are available to us.

REFERENCE

Schaible, W.L. "Choosing Weights for Composite Estimators for Small Area Statistics". Proceedings of the American Statistical Association, Section on Survey Research Methods, 1978, pp. 741-746.

TABLE 1.
Number of facilities in the national sample, and number of national and supplemental facilities in Illinois, Massachusetts, New York and Texas, by sampling stratum: 1977 NNHS

Stratum	Number of Sample Facilities												
	In total National Sample	Illinois			Massachusetts			New York			Texas		
		Total	Nat. Sample	Sup. Sample	Total	Nat. Sample	Sup. Sample	Total	Nat. Sample	Sup. Sample	Total	Nat. Sample	Sup. Sample
All Strata	1,698	143	100	43	142	66	76	148	116	32	155	102	53
Nursing Care Homes	1,292	94	77	17	117	59	58	77	77	0	128	90	38
< 15 beds	11	0	0	0	1	0	1	0	0	0	0	0	0
15-24 beds	30	1	1	0	5	3	2	1	1	0	0	0	0
25-49 beds	158	10	9	1	22	13	9	7	7	0	9	7	2
50-99 beds	398	25	19	6	33	16	17	12	12	0	44	30	14
100-199 beds	477	36	27	9	51	23	28	24	24	0	62	42	20
200-299 beds	134	15	14	1	3	2	1	19	19	0	10	8	2
300-599 beds	67	7	7	0	1	1	0	11	11	0	3	3	0
600+ beds	17	0	0	0	1	1	0	3	3	0	0	0	0
All Other Homes	318	49	23	26	25	7	18	71	39	32	27	12	15
< 15 beds	35	2	1	1	2	1	1	3	1	2	1	0	1
15-24 beds	17	1	0	1	3	1	2	2	1	1	1	1	0
25-49 beds	34	4	2	2	4	1	3	5	3	2	1	0	1
50-99 beds	60	8	3	5	4	1	3	6	3	3	5	2	3
100-199 beds	91	17	7	10	10	3	7	18	8	10	14	6	8
200-299 beds	34	10	5	5	1	0	1	13	8	5	4	2	2
300-599 beds	31	7	5	2	0	0	0	11	7	4	1	1	0
600+ beds	16	0	0	0	1	0	1	13	8	5	0	0	0
Unknown Type of Care Homes	88	No Supplemental Facilities Selected from Unknown Type of Care Homes											
< 15 beds	3												
15-24 beds	4												
25-49 beds	12												
50-99 beds	16												
100-199 beds	40												
200-299 beds	8												
300-599 beds	5												
600+ beds	-												

TABLE 2. ORIGINAL AND ALTERNATIVE ESTIMATES OF THE NUMBER AND VARIANCE OF NURSING HOME RESIDENTS FOR SELECTED NURSING HOME CHARACTERISTICS: 1977 NNHS.

NO. OF RESIDENTS HAVING DIFFICULTY CONTROLLING BOWELS AND BLADDER

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	337000.19	339087.42	334193.87	302771.62
VAR.	9173.09	9712.81	10867.49	9762.36
STAT-ARD		0.01	-0.01	-0.10
VAR-ARD		0.06	0.18	0.06

NO. OF RESIDENTS HAVING NO DIFFICULTY IN SPEECH

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	983872.34	987515.42	986316.18	884644.04
VAR.	12358.96	13250.29	14811.27	15203.14
STAT-ARD		0.00	0.00	-0.10
VAR-ARD		0.04	0.15	0.09

NO. OF RESIDENTS IN PROPRIETARY NURSING HOMES

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	888784.93	903459.43	894891.21	809385.35
VAR.	17878.35	18505.42	20531.86	19464.12
STAT-ARD		0.02	0.01	-0.09
VAR-ARD		0.04	0.15	0.09

NO. OF RESIDENTS IN SKILLED NURSING FACILITIES CERTIFIED FOR BOTH MEDICARE AND MEDICAID

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	190332.01	185501.37	174677.33	170541.12
VAR.	12105.57	11593.19	11379.69	11312.07
STAT-ARD		-0.03	-0.08	-0.10
VAR-ARD		-0.04	-0.06	-0.07

NO. OF RESIDENTS WITH PRIMARY DIAGNOSIS OF HARDENING OF THE ARTERIES

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	264377.00	265789.78	259536.81	235637.65
VAR.	8116.59	7918.46	7959.62	7956.67
STAT-ARD		0.01	-0.02	-0.11
VAR-ARD		-0.02	-0.02	-0.02

NO. OF RESIDENTS WITH PRIMARY DIAGNOSIS OF EMPHYSEMA

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	8004.15	8054.69	8139.68	7020.55
VAR.	1256.31	1225.49	1268.33	1177.60
STAT-ARD		0.01	0.02	-0.12
VAR-ARD		-0.02	0.01	-0.06

NO. OF RESIDENTS REQUIRING ASSISTANCE IN EATING

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	424907.81	424648.79	416136.23	374800.82
VAR.	9810.52	10234.85	11364.86	10463.72
STAT-ARD		0.00	-0.02	-0.12
VAR.ARD		0.04	0.16	0.07

NO. OF RESIDENTS OVER 85 YEARS OF AGE

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	449931.83	445185.41	430823.10	395922.41
VAR.	9210.48	9191.24	9256.51	9603.06
STAT-ARD		-0.01	-0.04	-0.12
VAR.ARD		-0.00	0.00	0.04

NO. OF RESIDENTS WHOSE PRIMARY REASON FOR BEING INSTITUTIONALIZED IS MENTAL ILLNESS

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	14398.06	14126.56	13944.30	12154.45
VAR.	1736.08	1645.50	1643.66	1568.64
STAT-ARD		-0.02	-0.03	-0.16
VAR.ARD		-0.05	-0.05	-0.10

NO. OF RESIDENTS RECEIVING THERAPY SERVICES FROM A LICENSED, REGISTERED OR PROFESSIONALLY TRAINED THERAPIST

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	847140.27	853348.87	840713.63	768795.44
VAR.	14821.96	15941.16	17630.15	16949.45
STAT-ARD		0.01	-0.01	-0.09
VAR.ARD		0.08	0.19	0.14

NO. OF RESIDENTS WITH PRIMARY SOURCE OF PAYMENT MEDICAID INTERMEDIATE CARE

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	362611.85	365610.27	361656.10	313400.74
VAR.	11844.01	11519.33	12041.50	11446.40
STAT-ARD		0.01	-0.00	-0.14
VAR.ARD		-0.03	0.02	-0.03

NO. OF RESIDENTS USING HEARING AIDS

	ORIGINAL	EST 1	EST 2	EST 3
STAT.	75930.95	74775.64	72821.84	66893.18
VAR.	4389.03	4169.49	4174.56	4059.76
STAT-ARD		-0.02	-0.04	-0.12
VAR.ARD		-0.05	-0.05	-0.08